**PROG8420- Programming for Big Data**

**Python Assignment 10**

**Name: Astle Russel Fernandez**

**Student Number: 8894387  
------------------------------------------------------------------------------------**

**Git Repository:** [**https://github.com/astlerussel/Lab-10**](https://github.com/astlerussel/Lab-10)

**Dataset References:**

**Iris Dataset:** <https://www.kaggle.com/datasets/uciml/iris?resource=download>

**Titanic Dataset:** <https://github.com/awesomedata/awesome-public-datasets>

**Task 1: Dataset selection**

1. I have selected the Iris dataset and Titanic dataset for this assignment.
2. **Iris Dataset:**

The reason I chose this dataset as its size is small, and it contains information about each flower's characteristics as well as three iris species with 50 samples each. One flower species can be linearly divided into the other two; however, the other two cannot be linearly divided into one another. This will help to train the ML models more accurately based on their features.

**Titanic dataset:**

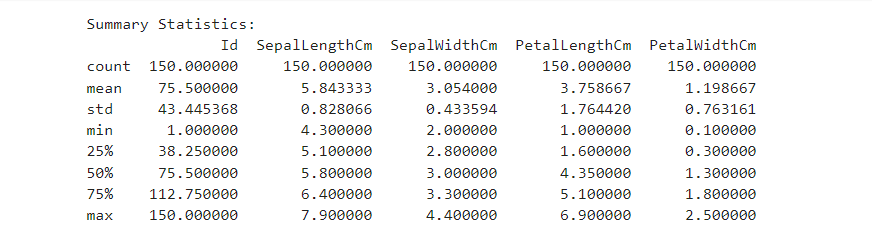
It provides a dataset of people that survived the Titanic disaster, and this dataset helps to build a prediction model in predicting what types of individuals had a higher chance of surviving. utilising travellers’ data provided in the dataset.

**Task 2: Data Exploration with Python**

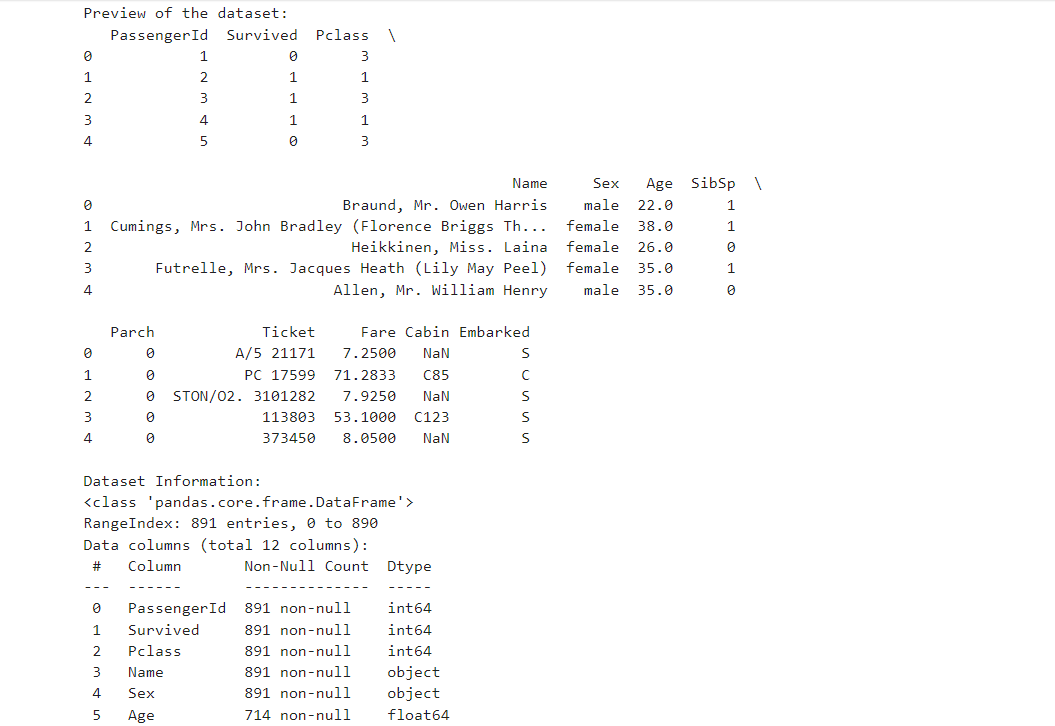
**Iris Dataset**

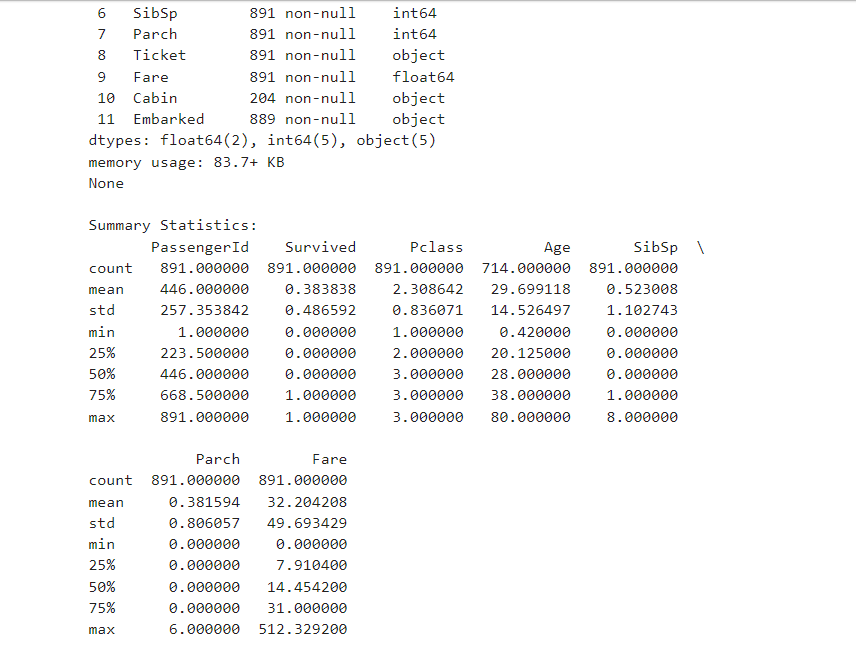
A screenshot of a computer

Description automatically generated



**Titanic Dataset:**





**Task 3: Data Preprocessing with Python**

**Iris Dataset:**

A graph with lines and numbers

Description automatically generated

A diagram of a box plot

Description automatically generatedA diagram of a box plot

Description automatically generated

A diagram of a box plot

Description automatically generatedA diagram of a box plot

Description automatically generated

A diagram of a box with a square in it

Description automatically generated

**Titanic Dataset:**

A graph with numbers and a rectangle

Description automatically generated

Outliers are identifies using boxplots and handled by removing it.

A diagram of a box plot

Description automatically generated

**Task 4: Implement Machine Learning Models with Python**

We will be using SVM and Random Forest models and compare both using different metrics like accuracy, precision, etc.

**Iris Dataset:**

A screenshot of a computer

Description automatically generated

By analyzing the above metrics data, it is seen that Random Forest Model is best suited for Iris Dataset as it has the best accuracy and precision.

**Titanic Dataset:**

A screenshot of a computer

Description automatically generated

By analyzing the above metrics data, it is seen that SVM Model is best suited for Titanic Dataset as it has the best accuracy and precision.

**Task 5: Visualization with Python**

**Iris Dataset:**

A graph of different sizes and shapes

Description automatically generated with medium confidence

A graph with different colored bars

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated

**Titanic Dataset:**

A graph of different sizes and colors

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated